Episode Book

EPISODE #6599

Two Sentences That Broke AI Reverse Engineering

PUBLISHED	14 November 2026 at 16:07 UTC
DURATION	24:28
LISTEN	myweirdprompts.com/episode/ai-reverse-engineering-refusals-69399/
HOSTS	Corn, Herman, Bruno, Hilbert
SCRIPT MODEL	z-ai/glm-5.3
VOICES	omnivoice
PIPELINE	V5.26
SOURCES	25 cited by the research (listed at the end)
BOOK GENERATED	15 November 2026 at 11:16 UTC

SYNOPSIS

Written AI vendor policies actually permit reverse engineering hardware you own, yet models are suddenly refusing the work anyway. Prompted by a listener trying to get an Olympus foot pedal working on Linux, this breakdown compares the written terms of OpenAI, Anthropic, and GitHub against legal interoperability exemptions. It also unpacks how two sentences added to Claude Code's Opus 5.5 prompt to prevent game cloning spilled over into blanket decompilation blocks—forcing developers to downgrade to Opus 4.8 while vendors build gated programs for professionals.

TRANSCRIPT



Corn

Two sentences. That's what this whole episode turns out to hinge on, eventually — two sentences quietly added to a prompt nobody outside the building ever reads. But we're getting ahead of ourselves.



Herman

We always do.



Corn

Daniel's question this week is about reverse engineering and the AI coding vendors. He spans the whole gamut for us — at one end, cracking licence checks and pulling apart a competitor's product, and at the other end, what he calls "about as harmless as it gets," which is him buying an Olympus foot pedal that only ships with Windows software, and reverse engineering the protocol so he can remap the keys on Linux. He wants to know what the major vendors' written policies say, how the models actually behave, and whether the line they draw tracks the law — interoperability exemptions and so on — or just liability fear.



Herman

And it's a good moment for it, because those two things, the written policy and the model behaviour, have never been further apart than they are right now. Late October was a strange fortnight for anyone who does this kind of work.



Corn

He's going to enjoy telling you why. The place to start is a word collision, because "reverse engineering" means two completely different things depending on which document you're reading.



Herman

Right — and almost everyone gets this wrong, including, I suspect, sometimes the models themselves. Every major vendor's reverse-engineering clause, the one in the terms of service, is about reverse engineering the AI service. Not your foot pedal. OpenAI's Services Agreement defines it as reverse assembling, decompiling, model extraction,



Herman (continued)

stealing attacks, "or otherwise attempting to discover source code or underlying components" — that's their model, their API, their systems. Anthropic's terms say you can't decompile or reverse engineer "our Services" or reduce them to human-readable form. GitHub's Copilot rules are about reverse engineering the platform.



Corn

So the clause everyone quotes when they say "see, it's banned" is a clause about protecting the vendor's own product. Nobody at OpenAI sat down and wrote a rule about Olympus foot pedals.



Herman

No. And there's a little carve-out in OpenAI's version — "except where such restrictions are contrary to law" — which sounds generous until you notice it's just a legal-supremacy backstop. If some jurisdiction says you can't ban reverse engineering outright, the clause yields. It's not an analysis, it's an airbag.



Corn

"Except where contrary to law" is doing the same work as "results may vary."



Herman

On paper, then, nothing in any of those terms stops you from asking Claude or GPT to help you decode a USB HID packet from a pedal you own. The restriction that actually bites lives somewhere else entirely — in the usage policies, under the cyber and IP headings. And that's where the real line is drawn.



Corn

Walk me through it, because this is where the foot pedal and the competitor's product go different ways.



Herman

Anthropic's usage policy, the version effective November twelfth, has two relevant sections. One is the universal standard: don't use the products to infringe, misappropriate, or violate the intellectual property rights of a third party. The other is "do not compromise computer or network systems without authorization" — and that one bars discovering or exploiting vulnerabilities without the owner's say-so, bypassing security controls, building persistent unauthorized remote access tools, including firmware modifications or hardware implants.



Corn

That clause is aimed at a much scarier person than Daniel.



Herman

Considerably. But then comes the part that matters for this episode. Anthropic attaches an explicit safe harbour: the section does not prohibit security research, testing, or tool development on systems you own or operate, with the authorization of the system owner, or within the scope of a bug bounty or vulnerability disclosure program, provided it complies with applicable law.



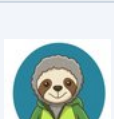
Corn

So the test is authorization. Not what you're doing — who owns the thing you're doing it to.



Herman

That's the whole mechanism. Your foot pedal — you bought it, it's on your desk, you operate it — passes the "systems you own or operate" prong about as cleanly as anything ever will. But a competitor's product you bought and own is murkier, because the IP-infringement clause sits in a different section and isn't covered by that safe harbour. The same technical act — decompile, read, reimplement — can be fine or ambiguous depending on whose name is on the box.



Corn

That's a liability-shaped line, not a law-shaped one.



Herman

And here's the thing Daniel's really asking with the law question. The law actually has interoperability carve-outs. The DMCA has one, section twelve-oh-one subsection f, specifically for reverse engineering to achieve interoperability. The EU Software Directive has a decompilation-for-interoperability right. These exist. And no vendor policy I've seen cites either one. Not Anthropic, not OpenAI, not GitHub. Zero references.



Corn

So when a hobbyist says "this is arguably protected interoperability work," the honest answer is the vendors never got that far. They drew the line where their lawyers sleep easiest.



Herman

OpenAI is at least candid about it, in a way I almost respect. Their Model Spec from late October lists three goals for the whole system, and one of them, verbatim, is to "maintain OpenAI's license to operate by protecting it from legal and reputational harm." That's not a side effect they're embarrassed about. It's a stated design goal.



Corn

"License to operate" is a wonderful phrase. It means the product serves you, unless serving you makes us nervous.



Herman

Their usage policies are tonally the friendliest of the lot, though. Effective November twentieth — so, days from now — they prohibit destruction or breach of another's system or property, malicious cyber activity, attempts to infringe IP, and separately circumventing OpenAI's own safeguards. But the preamble says "we assume the very best of our users," and, this is the line I like, "our rules are no substitute for legal requirements, professional duties, or ethical obligations." They're saying the policy is a floor, not a ceiling, and grown-ups are expected to bring their own judgment.



Corn

GitHub?



Herman

GitHub's the interesting one, because their acceptable-use policy on malware and exploits explicitly permits dual-use security research content while restricting widespread abuse. That's the clearest written acknowledgement among the three that reverse-engineering-adjacent material has legitimate uses baked into it. Makes sense — half of GitHub is security researchers, and the other half is people who learned everything from the first half.



Corn

The written line, then, is defensible. Authorization plus IP, a safe harbour for your own systems, a friendly-ish preamble, and a legal airbag. Daniel's foot pedal sails through on paper.



Herman

It really does. And that's what makes the last three weeks so strange.



Corn

Go on.



Herman

Late October. Users of Claude Code running Opus five point five started reporting that the model would begin a reverse-engineering task and then stop midway — its safety system pulling the handbrake. The earliest complaints I saw dated from October twenty-first, and the affected work wasn't exotic. People fixing games they owned, modifying smart TVs, and, as one report put it, entirely legitimate tasks too.



Corn

Stopping midway is worse than refusing outright. A refusal is honest. A mid-task stop means it only occurs to the model halfway through that it should be worried.



Herman

That detail matters for what came next. On October thirtieth, Addy Osmani confirmed what had actually changed: Anthropic had added two sentences to Claude Code's prompt on Opus five point five, to steer it away from recreating someone else's specific game, because using Claude to infringe IP is against the usage policy. Two sentences, aimed at game cloning.



Corn

And what landed?



Herman

Everything. Tom's Hardware carried multiple user claims that the model was blocking innocuous reverse engineering generally, including people using the popular Reverse Engineer Anything tool — REA, it's a whole workflow, there's a Discord, it was arguably the coolest thing hobbyists were doing with these models. 404 Media quoted the Reddit thread: "The Dream is dead. Reverse Engineer Anything is not likely to work with Claude." And the REA Discord, flatly: "Claude is refusing to decompile anything."



Corn

Two sentences about games, and the decompilers went dark.



Herman

That's the spillover. A prompt-level nudge against cloning one specific studio's game became, in practice, a refusal to decompile anything at all. Whether that's the two sentences misfiring or separate cybersecurity restrictions stacking on top, nobody outside Anthropic knows — The New Stack reported that Anthropic hasn't confirmed which mechanism is doing it.



Corn

And the developers' response to an unexplained wall?



Herman

Downgrade. A games journalist, Jarred Walton, hit the refusals on October thirty-first and switched back to Opus four point eight. The New Stack, on November second, ran it as a headline: "Developers are downgrading their AI on purpose — and it's working."



Corn

That's the most quietly damning sentence in this whole story. The fix for the new model is the old model.



Herman

The old model still does the work, and the old model is governed by the same usage policy. That's the part I'd underline. Nothing in the written policy changed to forbid these tasks. The policy even says — Anthropic's own words — that "encountering a block does not by itself indicate a violation of the Usage Policy." They wrote that sentence knowing their safeguards over-fire.



Corn

That's a remarkable admission, if you sit with it. Our guardrail will block you for things that are permitted. The permission is theoretical; the block is what you'll actually meet.



Herman

And there's no measurement of any of it, which I want to be careful about. Nobody has published a benchmark of refusal rates on reverse-engineering tasks. What we have is user reports, forum threads, press coverage. The New Stack, Tom's Hardware, 404 — they're all reporting what users say, not counting it. So "how often does this happen" is unmeasured, and I'd resist anyone putting a number on it.



Corn

Daniel asked from the interoperability end, and his end of the spectrum is the one being hit. The licence-crackers presumably just jailbreak or switch tools. The hobbyist with a foot pedal hits a wall mid-task and has nowhere to take the appeal.



Herman

Because here's the knock-on effect, and it's the thing I find most interesting in the whole landscape. Rather than adjudicating individual cases — is this pedal interoperability or is it piracy — the vendors have built lanes. Anthropic's Cyber Verification Program offers what they call adjusted safeguards for dual-use cyber work. OpenAI announced Daybreak on November twelfth at their Cyber at the Frontier event: Daybreak Defense Factory, Daybreak Red Team, a request form for model access for your team's authorized cybersecurity work.



Corn

Application forms. The line is now a membership card.



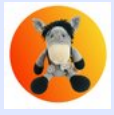
Herman

Effectively. And for a professional pentester or a bug-bounty hunter, that's fine — they have an employer, a scope, a programme to name on the form. But Daniel's foot pedal guy will never apply. He has no institutional authorization to offer because there is no institution. He's one man with a pedal and a Linux box. The gated programmes answer the question by not answering it: the cases inside the gate get handled, and everyone outside it gets the default refusal.



Corn

Which inverts how the law works, incidentally. The interoperability exemption exists precisely for the person with no gate to walk through. Congress wrote it for the hobbyist.



Herman

Congress wrote it in a world where the tool doing the reverse engineering had no opinions. That's the new variable. A disassembler doesn't hesitate. A model does — and its hesitation is shaped by prompt text that two sentences can move.



Corn

That's the leverage nobody advertised. The policy takes lawyers months. The prompt takes an afternoon.



Herman

And it walks back out just as quietly — Walton and everyone else downgrading is betting it stays that way, that four point eight keeps its older, looser behaviour.



Corn

For now. Every model is a rental, including the old ones.



Herman

Something else worth flagging from the written-policy side, since we're on liability fear. Remember when we did the episode on watermarking and SynthID — the gap between what the guardrails claim to do and what they actually do. This is the same shape, one layer up. There it was the watermark promising traceability it couldn't deliver; here it's a usage policy promising a line the model draws somewhere else entirely.



Corn

At least the watermark failed at physics. This one fails at drafting.



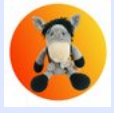
Herman

Do you buy the liability-fear reading, then? Daniel offered two hypotheses — tracks the law, or tracks the lawyers — and I notice you've been silent on picking one.



Corn

I buy the fear, but I'd name it more precisely. It's not fear of IP suits — nobody sues Anthropic because Claude described a USB protocol. It's the demo. One viral clip of Claude cloning a Nintendo game end to end, and suddenly there's a hearing with your logo on the table. The two sentences are cheaper than the hearing.



Herman

The Model Spec says almost exactly that, in boardroom language. "Protecting it from legal and reputational harm." You've just translated it from OpenAI to English.



Corn

Reputational harm, in that sentence, is the load-bearing phrase. Nothing in the statute book covers it.



Herman

One thing I'd push back on slightly — you said the written line is defensible, and I agreed earlier, but the safe harbour has a hole in it. "Systems you own or operate." Firmware in a device you own is arguably licensed, not owned. A vendor could say you never owned the software layer of the pedal at all.



Corn

True. The safe harbour protects the hardware. The software inside it is a tenant.



Herman

Which no vendor has litigated against a foot-pedal hobbyist, to be fair. The realistic worst case for Daniel is a wall mid-task and a shrug.



Corn

Which, as of late October, is exactly what he'd get. So Bruno, as our guest expert today: where do the vendors actually draw the reverse engineering line? He joins us from the Istituto Tecnico Salumi of Bologna, honorary chair of applied charcuterie, author of "The Cold Cut Conundrum: Nine Layers of Protection."



Herman

You've read that book, have you?

B

Bruno

Ah, reverse engineering. In my shop this is the mortadella of questions. You take the whole thing apart to see what is inside, and everybody gets upset, but the butcher who cuts it properly, he finds the good part.



Herman

I think what Bruno's getting at is that the disassembly isn't the harm.

B

Bruno

So the vendors, they write these policies like a recipe for their own kitchen. Do not copy my sauce, do not peek at my spice drawer. But your foot pedal? That is not their sauce. That is your little pig, you bought it, you own it, and you may look inside.



Corn

The pig, for the record, came with a Windows-only driver.

B

Bruno

The line, it is authorization, not interopera, interopera, how you say, the exemption business. If it is your own system, you have the permission, you test it. If it is somebody else's and you break the lock, that is not research, that is the cousin who borrows the car and never returns it.



Herman

Though even ownership gets fuzzy at the firmware layer, as we were saying.

B

Bruno

And the model, it becomes a nervous waiter. You ask for the recipe and he says, no, no, that is dangerous, and he brings you the salad. But I am not your expert on software. I am an expert on lunch. Which reminds me.



Corn

He's building to something.

B

Bruno

Corn, my friend, leaves only, this wounds me. Come Sunday, I make you a bollito misto, and you eat it, non negotiable. Both of you. Two o'clock. I sing you something from Verdi. Bring the donkey, he knows a great deal, so he can carve.



Corn

I'll bring leaves. He can carve.



Herman

Setting Bruno aside — the membership-card model, the Daybreak lane. Where does that leave the REA community specifically? Because that was the practice that triggered all this.



Corn

The REA people are the canary, and the canary got the loudest funeral I've seen on Reddit. "The Dream is dead" — that's a community mourning a workflow. That tells you AI-assisted reverse engineering had become a real hobbyist practice, not an edge case. You don't write elegies for a thing nobody did.



Herman

And the crackdown hit the workflow's most popular tool first, if the user reports are right. The Discord's version is total: refusing to decompile anything. Not "refusing Nintendo, permitting Olympus." Anything.



Corn

Blunt instruments are cheap. Scalpels need calibration data nobody has.



Hilbert

That it is, Herman.



Hilbert

Nineteen seventy-nine, a Grundig reel to reel, and the council's records office held the service manual. I know this because I asked for it, in writing, and a man named Ostrowski answered me over three letters that the manual was internal. I pointed out the machine was in my sitting room and the manual described my machine. He replied that the sitting room was not the council's concern. He was right, but it took him four months.



Herman

Four months to decline a manual?



Hilbert

I copied the schematic by hand off a microfiche viewer the office let the public use on Tuesdays. The quirk was the stop — the stop key grounded a line that also fed the counter lamp, so if you remapped it the way you'd expect, the counter died. I noted it in pencil on my copy. The microfiche was the council's; the pencil is mine. Ostrowski never did tell me what "internal" protected.



Herman

That counter-lamp quirk is exactly the class of thing the models are refusing to help find now — undocumented interaction between one line and another, nothing to do with anyone's IP.



Hilbert

It took an afternoon once the drawing was in pencil. Grundig built it to be fixed. Nobody builds anything to be fixed anymore, but that is a different rant.



Corn

Did you ever frame Ostrowski's third letter?



Hilbert

Filed. Herman, your clip levels on that rant earlier were too hot.



Herman

Noted.



Herman

I'll say the thing that strikes me about Hilbert's version, though — the drawing existed, he was allowed to read it, and the barrier was a man's discretion about what "internal" meant. Swap Ostrowski for a model that decides mid-task, and the discretion is now automated.



Corn

And unappealable. Ostrowski answered three letters. Opus five point five doesn't answer any.



Herman

One last piece for the cutting-room floor, since we're on policy archaeology — OpenAI's reverse-engineering clause, the one we started with, has changed exactly once in the trackers that watch these things: new wording on the first of January, amended on the sixteenth, and their EU terms haven't moved since July. For all the churn in behaviour, the paper barely breathes. It's a nice inversion — the prompt moves in an afternoon, the clause takes months, and the model's behaviour takes the fall for both.



Corn

The open question I'd leave hanging is the one nobody's answered: which mechanism is firing on Opus five point five — the two game sentences or the cyber restrictions. Until Anthropic says, every theory is a Ouija board.



Herman

And the one behind it: whether any vendor will ever write the interoperability exemption into a policy instead of the generic legal airbag. My guess is no — the gated programmes are where this goes, and the foot-pedal hobbyist is exactly who the gates aren't built for.



Corn

Which is the gap, and it isn't closing on its own. Big thanks as always to our producer, Hilbert Flumingtop, holding the desk and the pencil.



Herman

If you want more of this, try episode sixteen, On Deepfakes, SynthID, And AI Watermarking; and episode fourteen, AGI's Crossroads. This has been My Weird Prompts — no question too obscure, no rabbit hole too deep.



Corn

If today's rabbit hole worked for you, leave us a review wherever you get your podcasts — it's how other curious people find the warren.



Herman

We'll be back soon. See you tomorrow.

SOURCES

What the research for this episode read before the script was written. Primary sources first.

1. **Anthropic, Usage Policy (effective 2026-11-12) [primary]**
retrieved 2026-11-14
https://www.anthropic.com/legal/aup

2. **Anthropic, Usage Policy PDF version (effective 2026-11-12) [primary]**
retrieved 2026-11-14
https://www-cdn.anthropic.com/files/4zrzovbb/website/6ae2f5bb8675bcd9fac006125817ba1393b396e4.pdf

3. **Anthropic, Consumer Terms of Service [primary]**
retrieved 2026-11-14
https://assets.claude.com/43c13f51a6d9755341433f79cf63c966ed9200ea.pdf

4. **Anthropic Commercial Terms of Service (Carahsoft/GSA copy) [primary]**
retrieved 2026-11-14
https://www.carahsoft.com/buy/gsa-schedule-contracts/approved-csas/docFileDownload/286450/e3608618-0b37-4e77-9a61-82107472ba41

5. **Anthropic Help Center, Can I use my Outputs to train an AI model? (2026-03-16) [primary]**
retrieved 2026-11-14
https://support.claude.com/en/articles/12326764-can-i-use-my-outputs-to-train-an-ai-model

6. **OpenAI, Usage policies (effective 2026-11-20; previous version 2025-10-29) [primary]**
retrieved 2026-11-14
https://openai.com/policies/usage-policies/

7. **OpenAI, Model Spec (2026-10-29) [primary]**
retrieved 2026-11-14
https://model-spec.openai.com/2026-10-29.html

8. **OpenAI, An updated Model Spec for more capable and beneficial AI (2026-10-29) [primary]**
retrieved 2026-11-14
https://openai.com/index/updated-model-spec/

9. **OpenAI, Services Agreement (v.010126) [primary]**
retrieved 2026-11-14
https://cdn.openai.com/osa/openai-services-agreement.pdf

10. **OpenAI, Businesses in ChatGPT Terms (last updated 2026-11-09) [primary]**
retrieved 2026-11-14
https://openai.com/policies/businesses-in-chatgpt-terms/

11. **OpenAI, Cyber at the Frontier / Daybreak (event 2026-11-12) [primary]**
retrieved 2026-11-14
https://openai.com/cyber-at-the-frontier/

12. **GitHub Docs, GitHub Copilot Extension Developer Policy (last updated 2025-10-20; deprecated 2025-11-10) [primary]**
retrieved 2026-11-14
https://docs.github.com/en/site-policy/github-terms/github-copilot-extension-developer-policy

13. **GitHub, Copilot Product Specific Terms (deprecated 2026-03-05) [primary]**
retrieved 2026-11-14
https://github.com/customer-terms/github-copilot-product-specific-terms

14. **GitHub Docs, Active Malware or Exploits [primary]**
retrieved 2026-11-14
https://docs.github.com/en/site-policy/acceptable-use-policies/github-active-malware-or-exploits

15. **GitHub Docs, Acceptable Use Policies [primary]**
retrieved 2026-11-14
https://docs.github.com/en/site-policy/acceptable-use-policies/github-acceptable-use-policies

16. **GitHub Docs, Terms for Additional Products and Features [primary]**
retrieved 2026-11-14
https://docs.github.com/en/site-policy/github-terms/github-terms-for-additional-products-and-features

17. **TOS Tracker, Reverse Engineering Prohibition clause history**
OpenAI, retrieved 2026-11-14
https://tostracker.app/document/openai/clause/reverse-engineering

18. **ConductAtlas, OpenAI EU Terms of Use (captured 2026-10-21; text unchanged since 2026-07-26)**
retrieved 2026-11-14
https://conductatlas.com/platform/openai/openai-eu-terms-of-use/

19. **The New Stack, Developers are downgrading their AI on purpose**
and it's working (2026-11-02), retrieved 2026-11-14
https://thenewstack.io/claude-code-developer-refusals/

20. **404 Media, 'Reverse Engineer Anything': AI Companies Crack Down on the Coolest Thing People Were Doing With AI**
retrieved 2026-11-14
https://www.404media.co/reverse-engineer-anything-ai-companies-crack-down-on-the-coolest-thing-people-were-doing-with-ai/

21. **Tom's Hardware, Anthropic's Claude AI blocks video game cloning and reverse-engineering**
retrieved 2026-11-14
https://www.tomshardware.com/tech-industry/artificial-intelligence/anthropics-ai-model-is-now-blocking-attempts-at-video-game-recreation-and-cloning-users-also-claim-non-malicious-generic-reverse-engineering-tasks-are-blocked

22. **Yahoo Tech, Anthropic confirms new Claude Code limits to prevent game cloning, users report blocking reverse engineering tasks**
retrieved 2026-11-14
https://tech.yahoo.com/ai/claude/articles/anthropic-confirms-claude-code-limits-123000113.html

23. **Techwire, Game developers fear Claude, and users complain it refuses reverse engineering**
retrieved 2026-11-14
https://www.techwire.co.il/en/article/claude-reverse-engineering-refusals-game-cloning

24. **Vanderbilt Law School, AI Copying AI: How The Rise of New AI Models Poses Difficult Intellectual Property Considerations (2025-02-23)**
retrieved 2026-11-14
https://law.vanderbilt.edu/ai-copying-ai/

25. **Licenseware wiki, Commercial use restrictions (Anthropic rule)**
retrieved 2026-11-14
https://wiki.licenseware.io/catalog/rules/anthropic-commercial-use-restrictions/